COMPUTING THE ROOTED SUBTREE PRUNE AND DISTANCE IS NP-HARD

M Bordewich and C Semple

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2004/5 MARCH 2004

COMPUTING THE ROOTED SUBTREE PRUNE AND REGRAFT DISTANCE IS NP-HARD

MAGNUS BORDEWICH AND CHARLES SEMPLE

ABSTRACT. The graph-theoretic operation of rooted subtree prune and regraft is increasingly being used as a tool for understanding and modelling reticulation events in evolutionary biology. In this paper, we show that computing the rooted subtree prune and regraft distance between two rooted binary phylogenetic trees on the same label set is NP-hard. This resolves a longstanding open problem. Furthermore, we show that this distance is fixed parameter tractable when parameterised by the distance between the two trees.

1. INTRODUCTION

In evolutionary biology, it is becoming increasingly apparent that evolution is not necessarily tree-like because of reticulation events such as hybridisations and horizontal gene transfers. Consequently, the so-called "tree of life" is better represented as an acyclic digraph in which there is exactly one vertex that has in-degree zero and where the vertices of out-degree zero represent the present-day species.

One of the main tools used to understand and model reticulation events is a graph-theoretic operation called "rooted subtree prune and regraft". Loosely speaking, this operation prunes a subtree of a rooted tree and then reattaches this subtree to another part of the tree. The use of this tool in evolutionary biology dates back to at least 1990 [7] and has been regularly recognised since as an appropriate way to understand and represent reticulate evolution (for example, see [2, 10, 12, 15]). The reason for this is that two evolutionary (phylogenetic) trees describing the ancestral history of different genes for the same set of species may be inconsistent. However, for example, if this inconsistency can be resolved with a single hybridisation event, then one tree can be obtained from the other by a single rooted subtree prune and regraft operation.

Although reticulation events do occur in biology, they are still relatively rare and so a common problem is to find a representation of the available data which minimises the number of such events. This leads to the problem of computing the minimum number of rooted subtree prune and regraft operations required to

Date: 16 March 2004.

1991 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. Rooted phylogenetic tree, Rooted subtree prune and regraft.

The first author was supported by the New Zealand Institute of Mathematics and its Applications funded programme *Phylogenetic Genomics* and the second author was supported by the New Zealand Marsden Fund (UOC310).

FIGURE 1. Two rooted binary phylogenetic trees.

transform one evolutionary tree into another. In the context of these operations, this number is the "distance" between the two trees. It appears that this problem was first described in [7] and determining the complexity of computing this distance is stated as an open problem in several recent papers including [1, 12]. One of the two main results of this paper is to show that computing this distance is NP-hard. However, the other main result says that computing this distance is fixed parameter tractable when parameterised by the distance between the two trees. We describe these two results formally next.

A *rooted binary phylogenetic X-tree* is a rooted tree whose root has degree two and all other interior vertices have degree three, and whose leaf set is X. Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree and let $e = \{u, v\}$ be an edge of $\mathcal{T}$ where u is the vertex that is in the path from the root of $\mathcal{T}$ to v. Let $\mathcal{T}'$ be the rooted binary phylogenetic tree obtained from $\mathcal{T}$ by deleting e and then adjoining a new edge f between v and the component C_u that contains u in one of the following two ways:

(i) Creating a new vertex u' which subdivides an edge in C_u, and adjoining f between u' and v. Then, either suppressing the degree-two vertex u or, if u is the root of $\mathcal{T}$, deleting u and the edge incident with u, making the other end-vertex of this edge the new root.
(ii) Creating a new root vertex u' and a new edge between u' and the original root. Then adjoining f between u' and v and suppressing the degree-two vertex u.

We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a single *rooted subtree prune and regraft* (rSPR) operation. We define the rSPR *distance* between two arbitrary rooted binary phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, denoted $d_{\text{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)$, to be the minimum number of rooted subtree prune and regraft operations that is required to transform $\mathcal{T}_1$ into $\mathcal{T}_2$. It is well known that, for any such pair of trees, one can always obtain one from the other by a sequence of single rSPR operations. Thus this distance is well defined.

In the literature, it is sometimes unclear whether (ii) is allowed in the definition of an rSPR operation. However, if this is not part of the definition, then the rSPR distance on the collection of all rooted binary phylogenetic X-trees is not a metric. To see this, consider the two rooted binary phylogenetic trees shown in Fig. 1. It is easily checked that if we were to omit (ii) in the definition, then the rSPR distance to get from $\mathcal{T}_1$ to $\mathcal{T}_2$ is one, but the rSPR distance to get from $\mathcal{T}_2$ to $\mathcal{T}_1$ is two.

Theorem 1.1 is the first main result of this paper.

Theorem 1.1. *Computing the* rSPR *distance between an arbitrary pair of rooted binary phylogenetic X-trees is NP-hard.*

Theorem 1.1 has an interesting past. It was first thought to be proved by Hein *et al.* [8]. However, Allen and Steel [1] showed that a crucial lemma (Lemma 6) in their paper is incorrect. Nevertheless, their result and, in particular, this lemma could be recovered for a related tree operation called "tree bisection and reconnection". The crux in proving Theorem 1.1 is to show that the approach in [8] can still be used for the rSPR operation after modification of their main definition and re-proof of this particular lemma.

The second main result of this paper shows that despite Theorem 1.1 the problem of computing the rSPR distance between two rooted binary phylogenetic X-trees is fixed parameter tractable. In particular, we have the following theorem.

Theorem 1.2. *Computing the* rSPR *distance between an arbitrary pair of rooted binary phylogenetic X-trees is fixed parameter tractable when parameterised by d_{rSPR}.*

The proof of Theorem 1.2 closely follows the approach of Allen and Steel [1].

The paper is organised is as follows. The remainder of this section contains some preliminaries and some informative remarks. Section 2 contains the proof of Theorem 1.1 and Section 3 contains the proof of Theorem 1.2. In Section 4 we present an application of our approach. The rooted subtree prune and regraft operation is one of several tree rearrangement operations that are used in phylogenetics, we discuss its connection with these operations in the last section. Unless otherwise stated, the notation and terminology in this paper follows [14].

Let $\mathcal{T}$ be a rooted binary phylogenetic X-tree with root ρ. The set X is called the *label set* of $\mathcal{T}$ and is denoted by $\mathcal{L}(\mathcal{T})$. Now let V be a subset of the vertex set of $\mathcal{T}$. We denote by $\mathcal{T}(V)$ the minimal rooted subtree of $\mathcal{T}$ that connects the elements in V. Furthermore, the *restriction* of $\mathcal{T}$ to V is the rooted phylogenetic tree that is obtained from $\mathcal{T}(V)$ by suppressing all non-root vertices of degree two. This restriction is denoted by $\mathcal{T}|V$.

Rooted subtree prune and regraft is one of several important tree rearrangement operations in phylogenetics that induce metrics on the space of rooted or unrooted binary phylogenetic X-trees. In addition to the *unrooted* analogue of rooted subtree prune and regraft, two other types that have been extensively studied are *nearest neighbour interchange* (introduced independently in [11] and [13]) and *tree bisection and reconnection*. We describe these next.

A *binary phylogenetic X-tree* is an unrooted tree whose interior vertices all have degree three and whose leaf set is X. Let $\mathcal{T}$ be a binary phylogenetic X-tree and let $e = \{u, v\}$ be an edge of $\mathcal{T}$. Let $\mathcal{T}'$ be the binary phylogenetic X-tree that is obtained from $\mathcal{T}$ by deleting e, and then attaching the component C_v that contains v to the component C_u that contains u by adjoining a new edge f from C_v to

C_u so that, once degree-two vertices are suppressed, the resulting tree is a binary phylogenetic X-tree. The tree rearrangement operations that we now describe are restricted by how this new edge is adjoined. We begin with the least restrictive operation.

(i) We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *tree bisection and reconnection* (TBR) if there is no restriction on f.
(ii) We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by an (unrooted) *subtree prune and regraft* (uSPR) if one end-vertex of f is v.
(iii) We say that $\mathcal{T}'$ has been obtained from $\mathcal{T}$ by a *nearest neighbour interchange* (NNI) if one end-vertex of f is v and the other end-vertex subdivides an edge of C_u that is adjacent to an interior edge of $\mathcal{T}$ that is incident with v.

Analogous to rSPR, each $\Theta \in \{\text{NNI}, \text{uSPR}, \text{TBR}\}$ induces a metric on the space of binary phylogenetic X-trees. In particular, let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two binary phylogenetic X-trees. The Θ *distance* between $\mathcal{T}_1$ and $\mathcal{T}_2$, denoted $d_\Theta(\mathcal{T}_1, \mathcal{T}_2)$, is the minimum number of Θ operations that is required to transform $\mathcal{T}_1$ into $\mathcal{T}_2$. Again, it is well known that, for each Θ, one can always get from $\mathcal{T}_1$ to $\mathcal{T}_2$ by such a sequence of operations.

It is known that computing the NNI and TBR distances between two binary phylogenetic X-trees is NP-hard ([4] and [1], respectively). In this paper, we show that computing the rSPR distance between two rooted binary phylogenetic X-trees is also NP-hard. However, it remains an open problem to determine the complexity of computing the uSPR distance between two binary phylogenetic X-trees. Given the result in this paper and the previous results, it seems very likely that this is also NP-hard. Further discussion of these tree operations and their relationship to rSPR appears in the last section.

2. NP-Completeness

In this section, we prove Theorem 1.1. We begin by revising the definition of *maximum agreement forest* of [8]. Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. For the purposes of the definition, we regard the root of both $\mathcal{T}$ and $\mathcal{T}'$ as a vertex ρ at the end of a pendant edge adjoined to the original root. Furthermore, we also regard ρ as part of the label set of $\mathcal{T}$ and $\mathcal{T}'$ (see Fig. 2). An *agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is a collection $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$, where $\mathcal{T}_\rho$ is a rooted tree and $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k$ are rooted binary phylogenetic trees such that the following properties are satisfied:

(i) The label sets $\mathcal{L}(\mathcal{T}_\rho), \mathcal{L}(\mathcal{T}_1), \ldots, \mathcal{L}(\mathcal{T}_k)$ partition $X \cup \{\rho\}$ and, in particular, $\rho \in \mathcal{L}(\mathcal{T}_\rho)$.
(ii) For all $i \in \{\rho, 1, 2, \ldots, k\}$, $\mathcal{T}_i \cong \mathcal{T}|\mathcal{L}(\mathcal{T}_i) \cong \mathcal{T}'|\mathcal{L}(\mathcal{T}_i)$.
(iii) The trees in $\{\mathcal{T}(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ and $\{\mathcal{T}'(\mathcal{L}_i) : i \in \{\rho, 1, 2, \ldots, k\}\}$ are vertex-disjoint rooted subtrees of $\mathcal{T}$ and $\mathcal{T}'$, respectively.

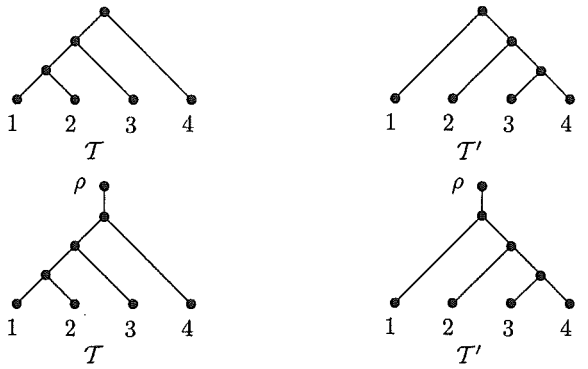

FIGURE 2. Two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$, pictured without (above) and with (below) their roots labelled.

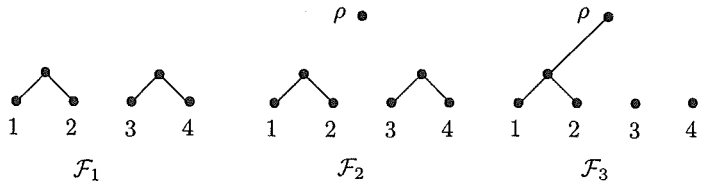

FIGURE 3. A maximum agreement forest $\mathcal{F}_1$ for $\mathcal{T}$ and $\mathcal{T}'$ of Fig. 2 under the definition in [8], and two maximum agreement forests $\mathcal{F}_2$ and $\mathcal{F}_3$ for $\mathcal{T}$ and $\mathcal{T}'$ under our definition.

A *maximum agreement forest* for $\mathcal{T}$ and $\mathcal{T}'$ is an agreement forest $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ in which k (the number of components minus one) is minimised. The minimum possible value for k is denoted by $m(\mathcal{T}, \mathcal{T}')$.

We remark here that the definition of maximum agreement forest given in [8] differs by the fact that ρ is not treated as part of the label set. Therefore, using the original definition, a maximum agreement forest for the two rooted binary phylogenetic trees $\mathcal{T}$ and $\mathcal{T}'$ shown in Fig. 2 consists of two components (for example, $\mathcal{F}_1$ in Fig. 3). However, for the definition given in this paper such a forest consists of three components (for example, $\mathcal{F}_2$ and $\mathcal{F}_3$ in Fig. 3). The crucial result in establishing Theorem 1.1 is Theorem 2.1.

Theorem 2.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees. Then $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = m(\mathcal{T}, \mathcal{T}')$.*

Proof. We first show that $m(\mathcal{T}, \mathcal{T}') \leq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. The proof of this inequality is by induction on $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. Assume that $d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}') = 1$. Let $\{A \cup \{\rho\}, B\}$ be the partition of $X \cup \{\rho\}$ induced by the "pruning" in an rSPR operation that transforms $\mathcal{T}$ into $\mathcal{T}'$. Then it is easily seen that an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$ is $\{\mathcal{T}|(A \cup \{\rho\}), \mathcal{T}|B\}$. Therefore, in this case, $m(\mathcal{T}, \mathcal{T}') \leq d_{\mathrm{rSPR}}(\mathcal{T}, \mathcal{T}')$. Now assume that the inequality holds for all rooted binary phylogenetic X-trees whose rSPR

distance is at most k. Suppose that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')=k+1$. Then there exists a rooted binary phylogenetic X-tree $\mathcal{T}''$ such that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}'')=k$ and $d_{\mathrm{rSPR}}(\mathcal{T}'',\mathcal{T}')=1$. By the inductive hypothesis, there is an agreement forest $\{\mathcal{T}_\rho,\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_k\}$ for $\mathcal{T}$ and $\mathcal{T}''$, and an agreement forest $\{\mathcal{T}'_\rho,\mathcal{T}'_1\}$ for $\mathcal{T}''$ and $\mathcal{T}'$. The partition $\{\mathcal{L}(\mathcal{T}'_\rho),\mathcal{L}(\mathcal{T}'_1)\}$ identifies a unique edge in $\mathcal{T}''$. Hence there can be at most one $i\in\{\rho,1,2,\ldots,k\}$ such that $\mathcal{L}(\mathcal{T}_i)\cap\mathcal{L}(\mathcal{T}'_\rho)\neq\emptyset$ and $\mathcal{L}(\mathcal{T}_i)\cap\mathcal{L}(\mathcal{T}'_1)\neq\emptyset$ (for otherwise, the induced subtrees in $\mathcal{T}''$ would not be disjoint, and $\{\mathcal{T}_\rho,\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_k\}$ would not be an agreement forest for $\mathcal{T}$ and $\mathcal{T}''$). If there is no such $i\in\{\rho,1,2,\ldots,k\}$, then $\{\mathcal{T}_\rho,\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_k\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. On the other hand, if there exists an $i\in\{\rho,1,2,\ldots,k\}$ such that $\mathcal{L}(\mathcal{T}_i)\cap\mathcal{L}(\mathcal{T}'_\rho)=\mathcal{L}_{i,\rho}\neq\emptyset$ and $\mathcal{L}(\mathcal{T}_i)\cap\mathcal{L}(\mathcal{T}'_1)=\mathcal{L}_{i,1}\neq\emptyset$, then $\{\mathcal{T}_j : j\in\{\rho,1,\ldots,k\}-\{i\}\}\cup\{\mathcal{T}_i|\mathcal{L}_{i,\rho},\mathcal{T}_i|\mathcal{L}_{i,1}\}$ is an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. In either case, we have $m(\mathcal{T},\mathcal{T}')\leq d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$.

We complete the proof by showing that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')\leq m(\mathcal{T},\mathcal{T}')$. We do this using induction on $m(\mathcal{T},\mathcal{T}')$. First assume that $m(\mathcal{T},\mathcal{T}')=1$. Let $\{\mathcal{T}_\rho,\mathcal{T}_1\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then the rSPR operation which prunes the subtree $\mathcal{T}|\mathcal{L}(\mathcal{T}_1)$ from $\mathcal{T}$, and regrafts this subtree in the correct place for $\mathcal{T}'$ demonstrates that $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')\leq m(\mathcal{T},\mathcal{T}')$. Thus the inequality holds for $m(\mathcal{T},\mathcal{T}')=1$. Now assume the inequality holds for all pairs of rooted binary phylogenetic X-trees for which there is an agreement forest of at most $k+1$ components. Suppose that $m(\mathcal{T},\mathcal{T}')=k+1$. Let $\{\mathcal{T}_\rho,\mathcal{T}_1,\mathcal{T}_2,\ldots,\mathcal{T}_{k+1}\}$ be an agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. Then there exists some $i\in\{1,2,\ldots,k+1\}$ such that $\mathcal{T}_i$ can be pruned from the rest of $\mathcal{T}$ by deleting a single edge. Consider the path in $\mathcal{T}'$ from the root of $\mathcal{T}_i$ to ρ. Let v be the first vertex on this path such that $v\in\mathcal{T}_j$ for some $j\in\{\rho,1,2,\ldots,k+1\}-\{i\}$. Because of (iii) in the definition of an agreement forest, v identifies a unique such $\mathcal{T}_j$. Let $\mathcal{L}_{i,j}=\mathcal{L}(\mathcal{T}_i)\cup\mathcal{L}(\mathcal{T}_j)$. Let $\mathcal{T}''$ be the tree obtained from $\mathcal{T}$ by pruning $\mathcal{T}_i$, and regrafting this subtree so that $\mathcal{T}''|\mathcal{L}_{i,j}\cong\mathcal{T}'|\mathcal{L}_{i,j}$. Now $\{\mathcal{T}_l : l\in\{\rho,1,\ldots,k+1\}-\{i,j\}\}\cup\{\mathcal{T}''|\mathcal{L}_{i,j}\}$ is an agreement forest for $\mathcal{T}''$ and $\mathcal{T}'$. Hence, by the inductive hypothesis, $d_{\mathrm{rSPR}}(\mathcal{T}'',\mathcal{T}')\leq k$ and, since $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}'')=1$, we have $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')\leq k+1=m(\mathcal{T},\mathcal{T}')$. □

With the revised definition of agreement forest and Theorem 2.1, the reduction of Theorem 8 in [8] can be applied to show that determining the rSPR distance between two rooted binary phylogenetic X-trees is NP-hard. To be precise, we define the decision problem rSPR as follows.

PROBLEM: rSPR
INSTANCE: Two rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$, and an integer k.
QUESTION: Is $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')\leq k$?

Theorem 1.1 is an immediate consequence of Corollary 2.2.

Corollary 2.2. *The decision problem* rSPR *is NP-complete.*

Proof. The reduction used in the proof of Theorem 8 in [8] can be applied using the revised definition of maximum agreement forest given above. This reduction is from "Exact Cover by 3-Sets (X3C)", see [6]. □

3. Fixed Parameter Tractability

In the previous section, we showed that computing the rSPR distance between two rooted binary phylogenetic X-trees is NP-hard. In spite of this, we now show that this problem is *fixed parameter tractable*, where we consider the rSPR distance itself to be the parameter. The idea behind fixed parameter complexity is that while the general case of computing rSPR distance is NP-hard, the cases in which one is generally interested, namely those in which the rSPR distance is small, may not be computationally infeasible. For example, we may be interested in comparing two evolutionary trees on a large number of species (> 1000) to determine how many hybridisation events must have occurred in order for the two trees to be consistent. Since hybridisation events are relatively rare, we would expect this number and, in particular, the rSPR distance between the two trees to be low (< 20). We show in this section that compared to the naive approach to computing the rSPR distance which takes time $O((2n)^{2k})$, the parameterised rSPR distance between two rooted binary phylogenetic X-trees may be computed in time $O(f(k)p(n))$, where $n = |X|$, k is the rSPR distance, f is some computable function, and p is a fixed polynomial. The importance of this result is in the separation of the variables n and k; it shows that, for a reasonable range of k, it may be possible to efficiently compute the rSPR distance even for trees with a very large leaf set. For further details on fixed parameter tractability, we refer the reader to [5]. It should be noted that while we have made the important theoretical step of establishing that rSPR is fixed parameter tractable, we have made no particular attempt to find the smallest function $f(k)$ possible.

We remark here that the authors of [4] and [1] have previously shown that computing the NNI and TBR distances between a pair of binary phylogenetic X-trees are fixed parameter tractable in their associated distances. It appears that the analogous result for uSPR is still open. The approach we take to establish the result for rSPR follows [1].

It is shown in [1] that two tree reduction rules could be used to reduce the size of the label set of a pair of (unrooted) binary phylogenetic X-trees while preserving the TBR distance between them. This process was shown to reduce any such pair of trees of TBR distance at most k apart to a new pair the same distance apart but on a label set of size linear in k. The tree reduction rules proposed were as follows.

Rule 1: Replace any pendant subtree that occurs identically in both trees by a single leaf with a new label.

Rule 2: Replace any chain of pendant subtrees that occurs identically in both trees by three new leaves with new labels correctly orientated to preserve the direction of the chain.

It is stated in [1] that Rule 1 preserves uSPR distance, and conjectured that Rule 2 also preserves uSPR distance. While this conjecture remains open, it is easily seen that rSPR distance is not preserved by the rooted analogue of Rule 2. To see this, let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the two rooted binary phylogenetic trees shown in

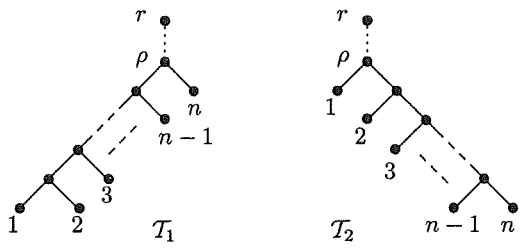

FIGURE 4. Two rooted binary phylogenetic trees.

Fig. 4, and let $\mathcal{T}_1'$ and $\mathcal{T}_2'$ be the binary phylogenetic trees obtained from $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, by adjoining a new leaf r to the root and viewing the resulting tree as unrooted (see Fig. 4, dotted lines). Although we can obtain $\mathcal{T}_2'$ from $\mathcal{T}_1'$ in a single uSPR operation by pruning r and regrafting it to the edge adjacent to 1, in the rooted setting this is not possible. By considering agreement forests, it is easy to see that, for n even, at least $n/2$ rSPR operations are needed to transform $\mathcal{T}_1$ into $\mathcal{T}_2$. This minimum $n/2$ can be achieved by taking each adjacent pair $2i-1, 2i$ $(i = 1 \ldots n/2)$ and the root to be the label sets of the $n/2 + 1$ components of the agreement forest.

Although Rule 2 does not preserve rSPR distance in the rooted setting, we will soon see that, together with Rule 1, the following modified version of Rule 2 does preserve rSPR distance.

> **Rule 2*:** Replace any chain of pendant subtrees that occurs identically and with the same orientation relative to the root in both trees by three new leaves with new labels correctly orientated to preserve the direction of the chain.

Rules 1 and 2* are illustrated in Figs. 5 and 6, respectively.

A *rooted abc-tree* is a rooted binary phylogenetic tree $\mathcal{T}$ whose label set includes the leaves a, b, c and has the following property. If v_a, v_b, v_c are the vertices adjacent to the leaves a, b, c, respectively, then $\{v_b, c\}$ are the two descendant neighbours of v_c and $\{v_a, b\}$ are the two descendant neighbours of v_b in $\mathcal{T}$. For example, $\mathcal{T}_1'$ and $\mathcal{T}_2'$ in Fig. 6 are rooted *abc*-trees.

The next lemma shows that if $\mathcal{T}$ and $\mathcal{T}'$ are both *abc*-trees on the same label set, then there is a maximum agreement forest in which a, b, and c are in the same component. Intuitively, this means there is a sequence of rSPR operations from $\mathcal{T}$ to $\mathcal{T}'$ which does not break up the common section containing a, b, and c. Once this lemma is established, it will then follow by a result analogous to Theorem 3.4 of [1] that an arbitrarily large number of additional pendant subtrees could be added to the path between v_a and v_c in both $\mathcal{T}$ and $\mathcal{T}'$ without changing the rSPR distance between them.

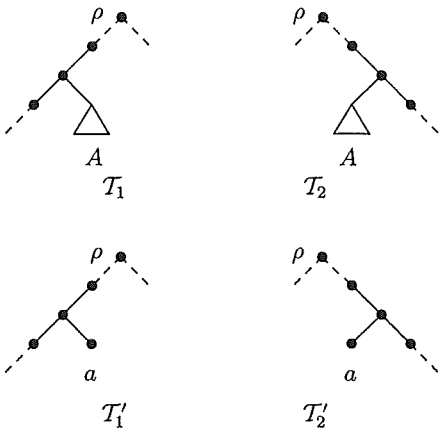

FIGURE 5. Two rooted binary phylogenetic trees reduced under Rule 1.

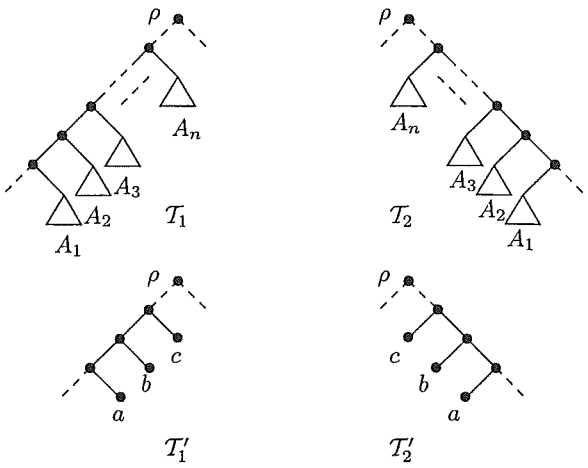

FIGURE 6. Two rooted binary phylogenetic trees reduced under Rule 2*.

Lemma 3.1. *If $\mathcal{T}$ and $\mathcal{T}'$ are two rooted abc-trees on the same label set, then there is a maximum agreement forest $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ for $\mathcal{T}$ and $\mathcal{T}'$ such that $a, b, c \in \mathcal{L}(\mathcal{T}_i)$, for some $i \in \{\rho, 1, 2, \ldots, k\}$.*

Proof. Let $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ be a maximum agreement forest for $\mathcal{T}$ and $\mathcal{T}'$. If there is some $i \in \{\rho, 1, 2, \ldots, k\}$ such that $a, b, c \in \mathcal{L}(\mathcal{T}_i)$, then we are done. Otherwise, let $\mathcal{L}_a$ be the set of descendant leaves of v_a in $\mathcal{T}$, not including a. Let $\mathcal{L}_c$ be the set of leaves that are not descendants of v_c in $\mathcal{T}$. Similarly, let $\mathcal{L}_a'$ be the set of descendant leaves of v_a in $\mathcal{T}'$, not including a, and let $\mathcal{L}_c'$ be the set of leaves that are not descendants of v_c in $\mathcal{T}'$. If there is such an i, let $i \in \{\rho, 1, 2, \ldots, k\}$

be such that both $\mathcal{L}(\mathcal{T}_i) \cap \mathcal{L}_a = \mathcal{L}_{i,a} \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_i) \cap \mathcal{L}_c = \mathcal{L}_{i,c} \neq \emptyset$. Lastly, if there is such a j, let $j \in \{\rho, 1, 2, \ldots, k\}$ such that both $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}'_a = \mathcal{L}'_{j,a} \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_j) \cap \mathcal{L}'_c = \mathcal{L}'_{j,c} \neq \emptyset$. There are six cases to consider:

(i) no such i and no such j exist,
(ii) $\exists$ i and no such j exists,
(iii) $\exists$ j and no such i exists,
(iv) $\exists$ i, j and $i \neq j$,
(v) $\exists$ i, j such that $i = j$ and $\mathcal{L}_{i,a} \cap \mathcal{L}'_{i,a} = \emptyset$, and
(vi) $\exists$ i, j such that $i = j$ and $\mathcal{L}_{i,a} \cap \mathcal{L}'_{i,a} \neq \emptyset$.

In each of these cases, we show that there is an agreement forest in which a, b, and c are in the same component, and this forest uses at most $k + 1$ components.

In case (i), either one of the leaves $x \in \{a, b, c\}$ is isolated in the agreement forest, or there is some r and some $s \in \{\rho, 1, 2, \ldots, k\}$ such that $a \in \mathcal{L}(\mathcal{T}_r)$ and $c \in \mathcal{L}(\mathcal{T}_s)$ (and b is either in $\mathcal{L}(\mathcal{T}_r)$ or $\mathcal{L}(\mathcal{T}_s)$). In the case that x is isolated, we may form an agreement forest of the same size (or smaller if more than one of a, b, c is isolated) by removing a, b and c from their respective trees and creating a new tree $\mathcal{T}|\{a, b, c\}$. In the case that there is no isolated leaf, we may form a smaller agreement forest by replacing $\mathcal{T}_r$ and $\mathcal{T}_s$ by $\mathcal{T}|(\mathcal{L}(\mathcal{T}_r) \cup \mathcal{L}(\mathcal{T}_s))$.

In cases (ii)-(vi), a, b and c must appear as isolated vertices in $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$. Otherwise, if $x \in \{a, b, c\}$ is in $\mathcal{L}(\mathcal{T}_i)$ say, then we may form an agreement forest of smaller size by replacing $\mathcal{T}_i$ and a, b, and c by $\mathcal{T}|(\mathcal{L}(\mathcal{T}_i) \cup \{a, b, c\})$. It is incorrectly claimed in [1] that, in their setting, if i exists and a, b, and c appear as isolated vertices, it is always possible to construct a smaller agreement forest thereby contradicting the maximality of $\{\mathcal{T}_\rho, \mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$. However, this is not always possible. Nevertheless, for each of the cases (ii)-(vi), there is an agreement forest of at most the same size in which a, b, and c are in the same component. In particular, such an agreement forest can be achieved by the following replacements:

(ii) Replace $\mathcal{T}_i$ and a, b, and c by $\mathcal{T}_i|\mathcal{L}_{i,a}$, $\mathcal{T}_i|\mathcal{L}_{i,c}$, and $\mathcal{T}|\{a, b, c\}$.
(iii) Replace $\mathcal{T}_j$ and a, b, and c by $\mathcal{T}_j|\mathcal{L}'_{j,a}$, $\mathcal{T}_j|\mathcal{L}'_{j,c}$, and $\mathcal{T}|\{a, b, c\}$.
(iv) Replace $\mathcal{T}_i$, $\mathcal{T}_j$, and a, b, and c by $\mathcal{T}_i|\mathcal{L}_{i,a}$, $\mathcal{T}_i|\mathcal{L}_{i,c}$, $\mathcal{T}_j|\mathcal{L}'_{j,a}$, $\mathcal{T}_j|\mathcal{L}'_{j,c}$, and $\mathcal{T}|\{a, b, c\}$.
(v) Replace $\mathcal{T}_i$ and a, b, and c by $\mathcal{T}_i|\mathcal{L}_{i,a}$, $\mathcal{T}_i|\mathcal{L}'_{i,a}$, $\mathcal{T}_i|(\mathcal{L}_{i,c} \cap \mathcal{L}'_{i,c})$, and $\mathcal{T}|\{a, b, c\}$.
(vi) Replace $\mathcal{T}_i$ and a, b, and c by $\mathcal{T}_i|(\mathcal{L}_{i,a} \cap \mathcal{L}'_{i,a})$, $\mathcal{T}_i|((\mathcal{L}_{i,a} \cup \mathcal{L}'_{i,a})) - (\mathcal{L}_{i,a} \cap \mathcal{L}'_{i,a}))$, $\mathcal{T}_i|(\mathcal{L}_{i,c} \cap \mathcal{L}'_{i,c})$, and $\mathcal{T}|\{a, b, c\}$.

It is easily checked that these are indeed agreement forests for $\mathcal{T}$ and $\mathcal{T}'$, and hence there exists a maximum agreement forest in which a, b and c appear in the same component. □

Proposition 3.2. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic X-trees. Let $\mathcal{T}'_1$ and $\mathcal{T}'_2$ be rooted binary phylogenetic X'-trees obtained from $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, by applying either Rule 1 or Rule 2*. Then $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) = d_{\mathrm{rSPR}}(\mathcal{T}'_1, \mathcal{T}'_2)$.*

Proof. The statement of this proposition is analogous to Theorem 3.4 in [1] which applies to Rule 1, Rule 2, and the TBR distance in the unrooted setting. The proof of the latter can be applied to the proof of this proposition using the definitions of rooted maximum agreement forest and rSPR operation given in this paper, and Lemma 3.1. □

We need one further lemma before we can tackle the main result of this section. Proposition 3.2 says that the tree reduction Rules 1 and 2* preserve rSPR distance; we now show that they can be repeatedly applied until the label set of the resulting rooted binary phylogenetic trees has size linear in the rSPR distance between them.

Lemma 3.3. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic X-trees. Let $\mathcal{T}_1'$ and $\mathcal{T}_2'$ be rooted binary phylogenetic X'-trees obtained from $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, by applying Rules 1 and 2* repeatedly until no further reduction is possible. Then $|X'| \leq 28 d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)$.*

Proof. By Theorem 2.1, $\mathcal{T}_1'$ and $\mathcal{T}_2'$ have a maximum agreement forest $\mathcal{S}_\rho, \mathcal{S}_1, \ldots, \mathcal{S}_k$ where $k = d_{\mathrm{rSPR}}(\mathcal{T}_1', \mathcal{T}_2')$. For $j = 1, 2$ and $i = \rho, 1, 2, \ldots, k$, let $n_j(i)$ denote the number of edges in $\mathcal{T}_j'$ which are incident with the subtree $\mathcal{T}_j'(\mathcal{L}(\mathcal{S}_i))$ if $i \neq \rho$ and let $n_j(i)$ denote one more than the number of edges incident with $\mathcal{T}_j'(\mathcal{L}(\mathcal{S}_\rho))$ if $i = \rho$. Then it follows from Lemma 3.7 of [1] that $|\mathcal{L}(\mathcal{S}_i)| \leq 7(n_1(i) + n_2(i))$ by simply substituting the rooted definition of maximum agreement forest into its proof. By Lemma 3.6 of [1] we have $\Sigma_{i \in \{\rho, 1, \ldots, k\}}(n_1(i) + n_2(i)) \leq 4k$, and so $|X'| = \Sigma_{i \in \{\rho, 1, \ldots, k\}} |\mathcal{L}(\mathcal{S}_i)| \leq 28k$. By Proposition 3.2, $d_{\mathrm{rSPR}}(\mathcal{T}_1', \mathcal{T}_2') = d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)$, and the result follows. □

We are now in a position to show that determining rSPR distance is fixed parameter tractable. Again, we formally deal with the decision problem.

Theorem 3.4. *The decision problem rSPR, parameterised by d_{rSPR}, is fixed parameter tractable.*

Proof. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic X-trees, and let k be an integer. Let $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) = d$. It follows by Lemma 3.1 of [1] that a pair of rooted binary phylogenetic X'-trees $\mathcal{T}_1'$ and $\mathcal{T}_2'$, obtained by applying Rules 1 and 2* repeatedly until no further reduction is possible, can be found in time polynomial in $|X|$ ($p(|X|)$ say). By Lemma 3.3, $|X'| \leq 28d$. If $|X'| \geq 28k$, we declare $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) > k$.

For a given rooted binary phylogenetic X-tree, there are $2|X| - 2$ edges that may be cut and at most $2|X| - 5$ to which a subtree may be regrafted to obtain a new rooted binary phylogenetic X-tree. Hence, for any such tree, there is at most $4|X|^2$ possible single rSPR operations. Therefore we can examine all possible paths from $\mathcal{T}_1'$ of length k in time $O((4|X'|^2)^k) = O((56k)^{2k})$. If one of these paths contains $\mathcal{T}_2'$, we declare $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) \leq k$, otherwise we declare $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) > k$. Hence we can answer the rSPR decision problem for $\mathcal{T}_1$ and $\mathcal{T}_2$ in time $O(f(k)p(|X|))$, where $f(k)$ is the computable function $(56k)^{2k}$ and $p(|X|)$ is the polynomial bound

for reducing the trees using Rules 1 and 2*. This satisfies the conditions for the decision problem rSPR to be fixed parameter tractable. □

Theorem 1.2 follows immediately from Theorem 3.4.

4. An Application of Maximum Agreement Forests

In this section we highlight a useful application of Theorem 2.1. The main result of this section, Theorem 4.1, is motivated by a question posed by Baroni and Steel [3].

A *cluster* C of a rooted binary phylogenetic X-tree $\mathcal{T}$ is a subset of X such that C is the set of label descendants of some vertex of $\mathcal{T}$. Confronted with finding the rSPR distance between two rooted binary phylogenetic X-trees, Theorem 4.1 says that one can "almost" break the problem into parts by considering common clusters between the two trees. Unfortunately, the theorem does not give an equality in doing this, but if one is only interested in a fast method that provides a good approximation, then this appears to be a reasonable approach.

Theorem 4.1. *Let $\mathcal{T}$ and $\mathcal{T}'$ be two rooted binary phylogenetic X-trees, and suppose that there is a cluster C common to $\mathcal{T}$ and $\mathcal{T}'$. Let $\overline{C} = X - C$. Then*

$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') - 1 \leq d_{\mathrm{rSPR}}(\mathcal{T}|C,\mathcal{T}'|C) + d_{\mathrm{rSPR}}(\mathcal{T}|\overline{C},\mathcal{T}'|\overline{C}) \leq d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}')$$

Proof. For the first inequality, let $\mathcal{F}_C = \{\mathcal{T}_{C,\rho}, \mathcal{T}_{C,1}, \ldots, \mathcal{T}_{C,k}\}$ be a maximum agreement forest for $\mathcal{T}|C$ and $\mathcal{T}'|C$, and $\mathcal{F}_{\overline{C}}$ be a maximum agreement forest for $\mathcal{T}|\overline{C}$ and $\mathcal{T}'|\overline{C}$. Then $\mathcal{F}_{\mathcal{T}} = \{\mathcal{T}_{C,\rho}|C, \mathcal{T}_{C,1}, \ldots, \mathcal{T}_{C,k}\} \cup \mathcal{F}_{\overline{C}}$ is an agreement forest of $\mathcal{T}$. And hence

$$\begin{aligned} d_{\mathrm{rSPR}}(\mathcal{T}|\overline{C},\mathcal{T}'|\overline{C}) + d_{\mathrm{rSPR}}(\mathcal{T}|C,\mathcal{T}'|C) &= |\mathcal{F}_C| - 1 + |\mathcal{F}_{\overline{C}}| - 1 \\ &= |\mathcal{F}_{\mathcal{T}}| - 1 - 1 \\ &\geq d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') - 1. \end{aligned}$$

For the second inequality, consider a maximum agreement forest $\mathcal{F}_{\mathcal{T}}$ for $\mathcal{T}$ and $\mathcal{T}'$. There are two cases to consider:

(i) there exists $\mathcal{T}_i \in \mathcal{F}_{\mathcal{T}}$ such that $\mathcal{L}(\mathcal{T}_i) \cap C \neq \emptyset$ and $\mathcal{L}(\mathcal{T}_i) \cap (\overline{C} \cup \{\rho\}) \neq \emptyset$, or
(ii) for all $\mathcal{T}_i \in \mathcal{F}_{\mathcal{T}}$, either $\mathcal{L}(\mathcal{T}_i) \subseteq C$ or $\mathcal{L}(\mathcal{T}_i) \subseteq (\overline{C} \cup \{\rho\})$.

In case (i), let $\mathcal{T}_{i,C}$ be the tree obtained from $\mathcal{T}_i|(C \cup x)$ by relabelling x as ρ, where $x \in \mathcal{L}(\mathcal{T}_i) \cap (\overline{C} \cup \{\rho\})$. Then $\mathcal{F}_C = \{\mathcal{T}_j \in \mathcal{F}_{\mathcal{T}} : \mathcal{L}(\mathcal{T}_j) \subseteq C)\} \cup \{\mathcal{T}_{i,C}\}$ is an agreement forest for $\mathcal{T}|C$ and $\mathcal{T}'|C$. Also $\mathcal{F}_{\overline{C}} = \{\mathcal{T}_j \in \mathcal{F}_{\mathcal{T}} : \mathcal{L}(\mathcal{T}_j) \subseteq \overline{C} \cup \{\rho\}\} \cup \{\mathcal{T}_i|(\overline{C} \cup \{\rho\})\}$

is an agreement forest for $\mathcal{T}|\overline{C}$ and $\mathcal{T}'|\overline{C}$. Hence

$$\begin{aligned} d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') &= |\mathcal{F}_{\mathcal{T}}| - 1 \\ &= |\mathcal{F}_C| + |\mathcal{F}_{\overline{C}}| - 1 - 1 \\ &\geq d_{\mathrm{rSPR}}(\mathcal{T}|C,\mathcal{T}'|C) + d_{\mathrm{rSPR}}(\mathcal{T}|\overline{C},\mathcal{T}'|\overline{C}). \end{aligned}$$

In case (ii), $\mathcal{F}_C = \{\mathcal{T}_i \in \mathcal{F}_{\mathcal{T}} : \mathcal{L}(\mathcal{T}_i) \subseteq C\} \cup \{\rho\}$ is an agreement forest for $\mathcal{T}_C$. Also $\mathcal{F}_{\overline{C}} = \{\mathcal{T}_i \in \mathcal{F}_{\mathcal{T}} : \mathcal{L}(\mathcal{T}_i) \subseteq \overline{C} \cup \{\rho\}\}$ is an agreement forest for $\mathcal{T}_{\overline{C}}$. Again

$$\begin{aligned} d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') &= |\mathcal{F}_{\mathcal{T}}| - 1 \\ &= |\mathcal{F}_C| + |\mathcal{F}_{\overline{C}}| - 1 - 1 \\ &\geq d_{\mathrm{rSPR}}(\mathcal{T}|C,\mathcal{T}'|C) + d_{\mathrm{rSPR}}(\mathcal{T}|\overline{C},\mathcal{T}'|\overline{C}), \end{aligned}$$

and hence the second inequality holds. This completes the proof of the theorem. □

We remark here that, for the NNI operation, it is shown in [9] that an analogous result to Theorem 4.1 is not possible: for any constant c, rooted binary phylogenetic X-trees $\mathcal{T}$ and $\mathcal{T}'$ can be constructed such that for some $C \subseteq X$ we have

$$d_{\mathrm{NNI}}(\mathcal{T},\mathcal{T}') \leq d_{\mathrm{NNI}}(\mathcal{T}|C,\mathcal{T}'|C) + d_{\mathrm{NNI}}(\mathcal{T}|\overline{C},\mathcal{T}'|\overline{C}) - c.$$

Also note that either of the inequalities in Theorem 4.1 can be tight: if $\mathcal{T} = \mathcal{T}'$, then for any cluster the second inequality is tight; if $\mathcal{T}'$ can be obtained from $\mathcal{T}$ by a single rSPR operation, then if C is the set of leaves in the pruned subtree the first inequality is tight.

5. Relating Distance Metrics on Phylogenetic Trees

In the last part of the introduction, we described three tree rearrangement operations associated with unrooted trees, namely, NNI, uSPR, and TBR. In addition to these operations for unrooted trees, each of NNI, uSPR, and TBR have rooted analogues. In the case of uSPR, we have already closely looked at its analogue. The rooted analogues of NNI and TBR applied to a rooted binary phylogenetic tree $\mathcal{T}$ are defined in the obvious way noting that the root of $\mathcal{T}$ can never be part of the pruned subtree.

For these tree rearrangement operations, it has been common practice to regard the unrooted and rooted cases as the same. However, while this is reasonable for NNI and TBR, it is not valid for SPR. The intuitive reason for this is that any single NNI operation performed in the unrooted (resp. rooted) setting can be performed with a single NNI operation in the rooted (resp. unrooted) setting. This also holds for TBR. As we have seen, this is not the case for SPR; a single operation in the unrooted setting in which the "root" itself is part of the pruned subtree cannot be performed by a single operation valid in the rooted setting. More precisely, we have the following proposition which compares the six operations.

Proposition 5.1. *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two rooted binary phylogenetic X-trees. Let $\mathcal{T}_1'$ and $\mathcal{T}_2'$ be the (unrooted) binary phylogenetic $(X \cup \{r\})$-trees obtained by attaching a pendant leaf r to the root of $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, and then regarding the resulting trees as unrooted. Then*

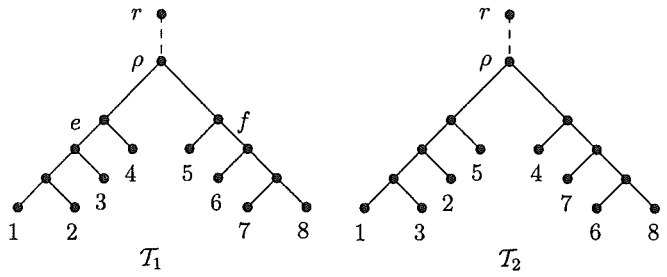

FIGURE 7. The distances between $\mathcal{T}_1$ and $\mathcal{T}_2$ differ for each metric.

(i) $d_{\mathrm{NNI}}(\mathcal{T}_1, \mathcal{T}_2) = d_{\mathrm{NNI}}(\mathcal{T}_1', \mathcal{T}_2')$.
(ii) $d_{\mathrm{TBR}}(\mathcal{T}_1, \mathcal{T}_2) = d_{\mathrm{TBR}}(\mathcal{T}_1', \mathcal{T}_2')$.
(iii) $d_{\mathrm{TBR}}(\mathcal{T}_1', \mathcal{T}_2') \leq d_{\mathrm{uSPR}}(\mathcal{T}_1', \mathcal{T}_2') \leq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) \leq d_{\mathrm{NNI}}(\mathcal{T}_1', \mathcal{T}_2')$.

Moreover, each of the inequalities in (iii) can be strict.

Proof. Part (i) can be checked in the following way. First consider an arbitrary unrooted NNI operation applied to $\mathcal{T}_1'$. If r is in the pruned subtree, there is another single NNI operation which does not have r in the pruned subtree and has the same result. Viewing the resulting tree as rooted, it is easily seen that this tree can therefore be obtained from $\mathcal{T}_1$ by a single rooted NNI operation. Furthermore, the analogous result for first applying an arbitrary rooted NNI operation to $\mathcal{T}_1$ also holds. Part (i) now follows.

Part (ii) can be obtained in the same way as Part (i).

By parts (i) and (ii), and the definitions of TBR, uSPR, and NNI, to prove (iii) it suffices to show that $d_{\mathrm{uSPR}}(\mathcal{T}_1', \mathcal{T}_2') \leq d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2)$. Consider a single rSPR operation applied to $\mathcal{T}_1$. It is easily checked that the resulting tree viewed as an unrooted tree can be obtained from $\mathcal{T}_1'$ by a single uSPR operation. Part (iii) now follows.

Lastly, it is not difficult to construct examples that show that each of the inequalities in (iii) can be strict (see Example 5.2). □

We end this section with an informative example.

Example 5.2. Consider Fig. 7. Viewing the two trees in this figure as their namesakes in the statement of Proposition 5.1, we show here that the inequalities in (iii) can all be strict simultaneously.

Using an exhaustive search, one can show $d_{\mathrm{TBR}}(\mathcal{T}_1', \mathcal{T}_2') = 2$, $d_{\mathrm{uSPR}}(\mathcal{T}_1', \mathcal{T}_2') = 3$, $d_{\mathrm{rSPR}}(\mathcal{T}_1, \mathcal{T}_2) = 4$, and $d_{\mathrm{NNI}}(\mathcal{T}_1', \mathcal{T}_2') = 5$. Furthermore, these values can be obtained as follows. For TBR, delete edges e and f, and reconnect the tree appropriately. For uSPR, first prune the subtree with leaves 1, 2, and 3, and regraft to the edge adjacent to 7, and then prune the subtree with leaves 4, 5, and r, and regraft to the

edge adjacent to 2. Lastly, prune the subtree with leaves 6, 7, and 8, and regraft to the edge adjacent to 4. For rSPR, we view the two trees as rooted and without the pendant edge with end-vertex r. In this case, prune 2 and regraft it to e, prune 7 and regraft it to f, and then prune 4 and 5 and swap their locations in two rSPR operations. Lastly, for NNI, follow the rSPR operations, but swapping 4 and 5 now requires three NNI operations.

References

[1] B. L. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Comb. 5 (2001) 1-13.

[2] M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, submitted.

[3] M. Baroni and M. Steel, Private communication.

[4] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang, On distances between phylogenetic trees, in: Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 1997, pp. 427-436.

[5] R. Downey, M. Fellows, Parameterized Complexity (Monographs in Computer Science), Springer Verlag, 1998.

[6] M. Garey, D. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.

[7] J. Hein, Reconstructing evolution of sequences subject to recombination using parsimony, Math. Biosci. 98 (1990) 185-200.

[8] J. Hein, T. Jing, L. Wang, K. Zhang, On the complexity of comparing evolutionary trees, Discrete Appl. Math. 71 (1996) 153-169.

[9] M. Li, J. Tromp, and L. Zhang, On the nearest neighbour interchange distance between evolutionary trees, J. Theor. Biol. 182 (1996) 463-467.

[10] W. Maddison, Gene trees in species trees, Syst. Biol. 46 (1997) 523-536.

[11] G. W. Moore, M. Goodman and J. Barnabas, An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posted by molecular data sets, J. Theor. Biol. 38 (1973) 423-457.

[12] L. Nakhleh, T. Warnow, and C. Randal Linder, Reconstructing reticulate evolution in species - theory and practice, in: Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB), 2004, in press.

[13] D. Robinson, Comparison of labelled trees with valency three, J. Combin. Theory 11 (1971) 105-119.

[14] C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.

[15] Y. Song and J. Hein, Parsimonious reconstruction of sequence evolution and haplotyde blocks: finding the minimum number of recombination events, in: Algorithms in Bioinformatics (WABI), G. Benson and R. Page, Eds., Lecture Notes in Bioinformatics, vol. 2812, 2003, pp. 287-302.

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: m.bordewich@math.canterbury.ac.nz, c.semple@math.canterbury.ac.nz